A Scalable Distributed Parallel Breadth-First Search Algorithm on BlueGene/L

2005 Gordon Bell Award - Finalist

Andy Yoo and Keith Henderson (CASC), Edmond Chow (DE Shaw Research), William McLendon (Sandia), Bruce Hendrickson, and Umit Catalyurek (Ohio State University)

Graph searches play an important role in analyzing large data sets since the relationship between data objects is often represented in the form of graphs, such as semantic graphs. Breadth-first searches (BFS) are of particular importance and are widely used in numerous applications to answer various user queries. A common query that arises in analyzing a semantic graph, for example, is to determine the nature of the relationship between two vertices in the graph. Such a query can be answered by finding the shortest path (or a set of paths meeting certain constraints) between those vertices using BFS. Further, BFS can be used for detecting communities, each of which is a set of strongly tied vertices, another important problem in semantic graph analysis. Searching very large graphs with billions of vertices and edges, however, poses challenges and calls for a distributed parallel BFS algorithm. The scalability of the distributed BFS algorithm for very large graphs, however, becomes a critical issue, since the demand for local memory and inter-processor communication increases as the graph size increases. In this work, we propose a scalable and efficient distributed BFS scheme that is capable of handling graphs with billions of vertices and edges and demonstrate its scalability on IBM BlueGene/L.

Team members (L-R) Keith Henderson (CAR), Umi Catalyurek (Ohio State University), and Andy Yoo (CAR), with Bill Gropp (ANL). Not pictured, Edmond Chow (DE Shaw Research), and Bill McLendon and Bruce Hendrickson (Sandia National Laboratory)

We achieve high scalability through a set of innovative optimization techniques. First, two-dimensional (2D) edge partitioning, also called 2D checkerboard partitioning, is used instead of the more conventional one-dimensional (1D) vertex partitioning to partition graphs. In the 2D edge partitioning, a level-expansion of the BFS is done by performing a column- and a row-wise communications. With the 2D partitioning, the number of processes involved in the collective communications is $O(\sqrt{P})$ in contrast to $O(P)$ of 1D partitioning, where P is the number of processors, and hence, we can reduce the communication time significantly.

In a distributed BFS algorithm, message buffers used in collective communications becomes a bottleneck for scalability since the buffer size increases as the number of processors increases. We attempt to relieve the memory bottleneck by controlling the size of message buffers. For this, we derive the bounds on the length of messages for Poisson random graphs. We show that given a random graph with n vertices, the expected message length is $O(n/P)$. That is, the length of messages sent out from a processor is bounded by the local problem size for Poisson random graphs. This optimization allows us to manage the local memory more efficiently and improve the scalability.

Finally, we have developed scalable collectives based on ring communications optimized for BlueGene/L. Here, taking advantage of BlueGene/L's high-bandwidth torus interconnect, we attempt to reduce the length of the ring by performing the ring communications in parallel. In the implementation, we explore the use of reduce-scatter (where the reduction operation is set-union) rather than straightforward use of all-to-all. It is shown that the reduce-scatter implementation significantly reduces message volume.

Our BFS scheme exhibits good scalability as it scales to a graph with 3.2 billion vertices and 32 billion edges on a BlueGene/L system with 32,768 nodes. To the best of our knowledge, this is the largest explicitly formed graph ever explored by a distributed algorithm. The performance characteristics of the proposed BFS algorithm are also analyzed and reported.

UCRL-MI-218107